Lecture 25: Model-agnostic regression
(i.e., how to do away with all our assumptions except iid)

1 The problem

So far in class we have mostly relied on 4 main assumptions:

- iid observations: $(X_1, Y_1), ..., (X_n, Y_n)$ are an iid sample
- linearity: $\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 X_1 + ... + \beta_d X_d = X^{\mathrm{T}}\beta$
- homoskedasticity: $\mathrm{var}(Y \mid X = x) = \sigma^2$
- normality: $Y \mid X = x \sim N(X^{\mathrm{T}}\beta, \sigma^2)$

In practice these will very often be suspect, especially the last three: regression functions can be very complex and nonlinear, variances may vary wildly, and normal distributions are probably the exception rather than the rule.

- Does this mean everything we learned so far is useless?

Well, even if everything we learned relied heavily on these assumptions, the tools and techniques picked up along the way would still be useful for less restrictive methods...

However, it turns out a lot of what we've done can still apply (with some slight modification) even if we drop the last 3 assumptions!

- the iid assumption can be weakened but is hard to drop entirely (otherwise our dataset is really just a single observation, and it's hard to say anything with $n = 1$)

2 Dropping homoskedasticity & normality

First, we will discuss dropping the constant variance and distributional assumptions together (but leaving linearity), since the work-around for both of these is basically the same.

Let's consider the usual least squares estimator from class

$$\widehat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

Recall we already showed this estimator is unbiased without using the constant variance or Gaussian assumptions. Let's see what the distributional properties are (before we required constant variance & Gaussian assumptions for this).

It will be useful to write the least squares estimator in a slightly different notation, where we let $X = (1, X_1, ..., X_d) \in \mathbb{R}^q = \mathbb{R}^{d+1}$ (note this is different from the design matrix) so that

$$\widehat{\beta} = \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\mathbb{P}_n(XY)$$

This follows since

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = n\mathbb{P}_n(XY) \quad \text{and} \quad \mathbf{X}^{\mathrm{T}}\mathbf{X} = n\mathbb{P}_n(XX^{\mathrm{T}})$$

where as usual $\mathbb{P}_n(Z) = \frac{1}{n}\sum_{i=1}^n Z_i$ denotes sample averages.

- note: you should check for yourself that the LHS and RHS above are equal

Before considering distributional properties let's first see what this estimator is converging to, as a sanity check.

Only assuming finite variances, the weak law of large numbers (with the continuous mapping theorem) tells us that

$$\widehat{\beta} = \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\mathbb{P}_n(XY) \;\to\; \mathbb{E}(XX^{\mathrm{T}})^{-1}\mathbb{E}(XY)$$

in probability. Let's think about what this large sample limit is. Note $Y = X^{\mathrm{T}}\beta + \epsilon$ where $\mathbb{E}(\epsilon \mid X) = 0$ (but we do not assume ϵ has constant variance or is normal). Therefore

$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}\{X(X^{\mathrm{T}}\beta + \epsilon)\} \\ &= \mathbb{E}(XX^{\mathrm{T}})\beta + \mathbb{E}(X\epsilon) \\ &= \mathbb{E}(XX^{\mathrm{T}})\beta\end{aligned}$$

where the last equality follows since

$$\mathbb{E}(X\epsilon) = \mathbb{E}\{X\mathbb{E}(\epsilon \mid X)\} = 0$$

Therefore

$$\widehat{\beta} \to \beta$$

in probability, only assuming iid and linearity. What about distributional properties?

Recall before (assuming constant variance and normality) we had conditioning on X that

$$\widehat{\beta} - \beta \sim N(0, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

or in our new notation,

$$\widehat{\beta} - \beta \sim N\left(0, \frac{\sigma^2}{n}\mathbb{P}_n(XX^{\mathrm{T}})^{-1}\right)$$

i.e.,

$$\left\{\left(\frac{\sigma^2}{n}\right)\mathbb{P}_n(XX^{\mathrm{T}})^{-1}\right\}^{-1/2}\left(\widehat{\beta} - \beta\right) \sim N\left(0, I\right).$$

Does a similar result hold without homoskedasticity/normality?

We have

$$\begin{aligned}\widehat{\beta} - \beta &= \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\mathbb{P}_n(XY) - \mathbb{E}(XX^{\mathrm{T}})^{-1}\mathbb{E}(XY) \\ &= \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\Big\{\mathbb{P}_n(XY) - \mathbb{E}(XY)\Big\} - \Big\{\mathbb{E}(XX^{\mathrm{T}})^{-1} - \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\Big\}\mathbb{E}(XY)\end{aligned}$$

where for the first term by Slutsky's theorem

$$\sqrt{n}\mathbb{P}_n(XX^{\mathrm{T}})^{-1}\Big\{\mathbb{P}_n(XY) - \mathbb{E}(XY)\Big\} \approx \sqrt{n}\mathbb{E}(XX^{\mathrm{T}})^{-1}\Big\{\mathbb{P}_n(XY) - \mathbb{E}(XY)\Big\}$$

and for the second term

$$\sqrt{n}\Big\{\mathbb{E}(XX^{\mathrm{T}})^{-1} - \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\Big\}\mathbb{E}(XY) = \sqrt{n}\mathbb{P}_n(XX^{\mathrm{T}})^{-1}\Big\{\mathbb{P}_n(XX^{\mathrm{T}}\beta) - \mathbb{E}(XX^{\mathrm{T}}\beta)\Big\}$$
$$\approx \sqrt{n}\mathbb{E}(XX^{\mathrm{T}})^{-1}\Big\{\mathbb{P}_n(XX^{\mathrm{T}}\beta) - \mathbb{E}(XX^{\mathrm{T}}\beta)\Big\}$$

where we again used Slutsky's in the last line. Therefore combining gives

$$\begin{aligned}\sqrt{n}(\widehat{\beta} - \beta) &\approx \sqrt{n}\mathbb{E}(XX^{\mathrm{T}})^{-1}\mathbb{P}_n\{X(Y - X^{\mathrm{T}}\beta)\} \\ &\rightsquigarrow N\left(0, \mathrm{var}\Big\{\mathbb{E}(XX^{\mathrm{T}})^{-1}X(Y - X^{\mathrm{T}}\beta)\Big\}\right) \\ &= N\left(0, \{\mathbb{E}(XX^{\mathrm{T}})^{-1}\}\,\mathrm{var}\Big\{X(Y - X^{\mathrm{T}}\beta)\Big\}\,\{\mathbb{E}(XX^{\mathrm{T}})^{-1}\}\right) \\ &\equiv N(0, BMB)\end{aligned}$$

i.e.,

$$\left(\frac{BMB}{n}\right)^{-1/2}(\widehat{\beta} - \beta) \rightsquigarrow N(0, I)$$

The matrix BMB is called the "sandwich variance" (can you see why?). When the assumptions of homoskedasticy+normality hold, BMB reduces to an analog of the variance we computed earlier in class, since in that case the meat equals $M = \sigma^2 B^{-1}$.

You can compute sandwich variance-based CIs in R with the `sandwich` package. I recommend using it whenever you use linear regression in practice - if the homoskedasticity+normality assumptions are actually correct, you should get the same results, and if they aren't correct, then you still have valid CIs!

3 Dropping linearity

There are two main ways we could drop the linearity assumption:

- change our target from "the regression function" to its best linear approximation
- target the regression function but change our estimator to accommodate nonlinearity

The second approach forms the basis of many flexible nonparametric tools in statistics and machine learning. We will come back to it in a few lectures. Today we consider the first approach.

3.1 The simple linear regression case

In fact we have used this approach before, when defining the target parameters

$$\beta_0 = \mathbb{E}(Y) - \beta_1\mathbb{E}(X)\ ,\ \beta_1 = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$$

in simple linear regression. Specifically we showed that these coefficients are those of the best linear predictor in the sense that

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\arg\min}\,\mathbb{E}[\{Y - (b_0 + b_1 X)\}^2]$$

without assuming *anything* about the form of the regression function $\mathbb{E}(Y \mid X)$.

3.2 Best linear approximation in multiple regression

In turns out a similar results holds in multiple regression.

Define

$$\beta = \underset{b \in \mathbb{R}^q}{\arg\min}\, \mathbb{E}\{(Y - b^{\mathrm{T}} X)^2\}$$

We can find an exact expression for this minimizer by differentiating the RHS, setting equal to zero, and solving:

$$\begin{aligned} \frac{\partial}{\partial b} \mathbb{E}\{(Y - b^{\mathrm{T}} X)^2\} &= \mathbb{E}\left\{ \frac{\partial}{\partial b} (Y - b^{\mathrm{T}} X)^2 \right\} \\ &= \mathbb{E}\{2X(Y - X^{\mathrm{T}} b)\} \end{aligned}$$

Setting to zero gives

$$\beta = \mathbb{E}(XX^{\mathrm{T}})^{-1} \mathbb{E}(XY) \tag{1}$$

which is the population version of our familiar least squares estimator

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbb{P}_n(XX^{\mathrm{T}})^{-1}\mathbb{P}_n(XY)$$

Therefore we can target β defined as the best linear predictor, without assuming the regression function is actually linear! And we can use our familiar least squares estimator.

- this is pretty nice.

3.3 Do we have to re-derive all our bias/variance, distributional, CI results?

In fact, all the work we just did in the heteroskedastic/non-normal case transfers over to this new otherwise model-agnostic parameter.

This follows since in the previous section we only ever used the fact that

$$\mathbb{E}\{X(Y - X^{\mathrm{T}}\beta)\} = 0$$

and not explicitly that $\mathbb{E}(Y \mid X) = X^{\mathrm{T}}\beta$ (go back and check for yourself). And the above holds for the best linear predictor version of β based on the result in (1). In other words the residuals $\epsilon = Y - X^{\mathrm{T}}\beta$ here do not have conditional mean zero, but we never actually required that, only that $\mathbb{E}(X\epsilon) = 0$.

Therefore we can simply re-interpret all our results based on estimating the *best linear approximation* of the regression function rather than the regression function itself!

- this shows that linear regression is surprisingly robust, only depending on an iid assumption (at least if we use the modified sandwich variance)

4 Conclusion

In practice, this means you can use exactly the same estimator we derived earlier, with a slightly modified std error estimator, and reap all the same benefits without any of the assumptions (except pesky iid)